Structural Genomic Variation Underlying Deer Speciation

Faezeh Azimi Chetabi[1], Aaron Shafer[1]

[1] 2140 East Bank Drive, Environmental and Life Sciences Graduate Program, Trent University, Peterborough, Ontario, Canada K9J 7B8

Correspondence: faezehazimichetabi@trentu.ca

Abstract

Speciation is a key driver of biodiversity and understanding its genomic underpinnings is important for predicting and managing biodiversity, especially in the face of environmental changes. Structural variants (SVs) are large-scale (>50 bp) changes in the genome and have been implicated in adaptive divergence and reproductive isolation. We investigated the role of SVs in the speciation and divergence of two deer species (*Odocoileus* spp.) across their North American range. Using multiple genome assemblies and a large short-read dataset, our bioinformatics workflow revealed SVs and genome features that were unique to each species. We found no difference in enhancer motifs found in SVs compared to the rest of the genome, while the predominance of species-specific SVs were deletions and insertions suggests that these variants may show higher likelihoods of fixation within populations. Further, while most SVs were intergenic, some genes found to be impacted by species-specific SVs were under positive selection inferred from dN/dS. The SV-affected genes were involved in reproduction, sensory adaptation, and metabolic pathways, which have clear links to fertility and deer biology, supporting their role in reproductive isolation and speciation.

Keywords: Structural variant, Short-read, Speciation, *Odocoileus hemionus*, *Odocoileus virginianus*

1. Introduction

Speciation is a fundamental process driven most notably by ecological divergence and geographic isolation (Sobel, 2010), with the ultimate mode(s) of divergence leaving tell-tell signals on the genome (Shang et al. 2023). An emerging body of research has identified genomic structural variation (SV) in the formation of reproductive barriers and speciation. Early work used traditional cytogenic analysis to detect SVs, with for example two distinct clades of Brocket deer (*Mazama* spp.) having karyotypic differences leading to reproductive isolation (Abril et al., 2010). With the advent of high-throughput sequencing and novel bioinformatics tools, assaying SVs at base-pair resolution along the genome has become tenable, opening a new line of inquiry as to the role of SVs in adaptation and ultimately speciation.

Structural variants are generally considered features >50bp and fall into several main types: deletions, duplications, inversions, insertions and translocations. Different sequencing and analytical techniques can be employed to detect SVs. Short-read data with paired read and split read analyses offer high-resolution breakpoint detection, though their effectiveness may depend on factors like insert size or coverage depth (Escaramís et al., 2015). Read depth analysis, often used with both short and long-read data, provides information into SV presence by examining read density, but is less precise with breakpoint resolution (Escaramís et al., 2015). For long-read data, de novo assembly and genome alignments are frequently employed, and offers the advantage of identifying novel SVs without relying on reference genomes, though it is more resource-intensive (Escaramís et al., 2015). Long-read sequencing is generally better than SR sequencing for detecting complex and long SVs because the latter struggles with detecting novel insertions and resolving complex rearrangements due to the limited read length (Mahmoud et al., 2019).

Structural variants such as inversions and translocations can play a role in promoting speciation through mechanisms like reproductive isolation (Homolka et al., 2007). One classical model suggests that underdominance in heterokaryotypes, where individuals with chromosomal rearrangements exhibit reduced fertility, could lead to hybrid sterility (Homolka et al., 2007). Autosomal translocations have also been shown to cause male sterility in laboratory mice due to disrupted synapsis and meiotic silencing

(Homolka et al., 2007). Similarly, Yang et al., (2024) showed convergent evolutionary signals of SVs associated with domestication, adaptation, and production traits; here deletions in key genes were linked to increased litter size in sheep, suggesting a novel neural regulatory mechanism influencing female reproduction traits (Yang et al., 2024). Also, Gompert et al. (2025) demonstrated that complex structural variants, including inverted translocations, can independently arise in separate populations and repeatedly drive adaptive divergence by suppressing recombination and clustering functionally important genes, revealing a predictable mechanism of local adaptation and evolutionary change. Traditional sequencing approaches focusing on outliers and protein-coding differences often fail to explain the evolution of complex traits, highlighting the importance of non-coding regions like enhancers (Kaplow et al., 2023). Structural variants can drive phenotypic diversity and speciation by influencing gene expression (Kaplow et al., 2023), with enhancers often connecting genetic changes to species-specific phenotypes (Kaplow et al., 2023). Collectively, these findings highlight SVs and enhancer regions as potential drivers of evolutionary divergence, shaping unique phenotypes and enabling evolution of traits across species.

The *Cervidae* (Deer) family offers an interesting case study for exploring SVs and genetic history. This family encompasses over 60 species with its origins dating back nearly 20 million years (Gonzalez and Duarte, 2020). The red brocket deer (*Mazama americana*) for example exhibits significant chromosomal diversity, with at least six distinct karyotypes across South America, suggesting the presence of multiple cryptic species (Cursino et al., 2014). Studies on hybrids between these karyotype lineages reveal post-zygotic reproductive isolation, with hybrids showing sterility or subfertility due to chromosomal imbalances impacting ovarian structure, oocyte development, and embryonic genome activation (Cursino et al., 2014). Two closely related species, white-tailed deer (WTD; *Odocoileus virginianus*) and mule deer (MD; *O. hemionus*), have a relatively recent speciation history, appearing to have split in allopatry around 1 million years ago (Ma) but experience contemporary hybridization (Kessler et al., 2023). Kessler et al., (2023) suggested potential incompatibilities given the lack of nuclear introgression and historical gene flow. Structural variations, such as inversions that suppress recombination in heterozygotes, might help

maintain distinct genetic clusters between two species (Stevison et al., 2011). Barriers to gene flow could further be reinforced by genetic incompatibilities caused by SVs, suggesting that SVs might have play a role in species divergence (e.g. Martin et al., 2017 & Baurens et al., 2019).

Here we used multiple genome assemblies, treating them as assembly data, and short-read (SR) sequencing data from mule deer and white-tailed deer across their range to identify unique SVs within species. By employing a range of detection tools and identifying commonalities, this study provides new insights into how SVs lead to genomic divergence, reproductive isolation, and the broader evolutionary processes shaping the genetic landscape of *Odocoileus* speciation.

2. Materials and Methods

We collated genomic short read datasets from 79 samples, comprising 57 for WTD and 22 for MD (PRJNA830519). We acquired three genome assemblies from the National Center for Biotechnology Information (NCBI) database. These included two genome datasets for white-tailed deer and one genome dataset for mule deer, each constructed from different types of sequencing data (Table 1). We employed two distinct approaches to detect SVs. First, we used three haploid genome datasets and treated them as assembly data for genome analysis section (Figure 1). Second, we used the short-read data in conjunction with highest quality genome per species to detect SVs for short rad analysis section (Figure 1).

2.1. Genome assemblies

We used the JAJQKH000000000 genome (WTD) as the primary reference for the genome assembly data analysis (Figure 1) due to it being the most contiguous assembly (Table 1). We aligned the two other deer assemblies to JAJQKH000000000 using Minimap2 (Li, 2018) with default settings (see Table 1; Figure 1). The generated SAM files were then converted, sorted, and indexed using SAMtools (Danecek et al., 2021). Subsequently, SVs were detected using both Delly (Rausch et al., 2012) and Sniffles (Sedlazeck et al., 2018) with default settings. We applied a minimum quality score threshold of 100 for the identified SVs using BCFtools (Danecek et al., 2021) to filter out low quality SVs. We then merged the outputs with

SURVIVOR (Jeffares et al., 2017), selecting SVs that were detected by both tools. This led to two genome comparison SV datasets referred to as White_SV and Mule_SV.

2.2. Short-read analysis

We removed adaptors from the raw fastq files using Trimmomatic (Bolger et al., 2014) with a MINLEN threshold of 36. We employed Bowtie2 (Langmead et al., 2012) to align all samples to both the reference genomes of white-tailed deer (JAJQKH000000000) and mule deer (JAJLRB000000000). This resulted in the following mapping conventions: mule2mule, mule2white, white2white, and white2mule (see Figure 1). Following this, we used SAMtools (Danecek et al., 2021) to convert, sort, and index the SAM files. To ensure data integrity, duplicate reads were eliminated using Picard (Broad Institute, 2018) with REMOVE_SEQUENCING_DUPLICATES set to TRUE.

We employed three different SV calling tools for SR data: Delly (Rausch et al., 2012), Manta (Chen et al., 2016), and Lumpy (Layer, 2014), all with default settings. Further refinement and comparative analysis of VCFs were performed using BCFtools (Danecek et al., 2021), applying a minimum quality score threshold of 100 to filter out low quality SVs. We merged the SV call sets from Delly, Manta, and Lumpy using SURVIVOR, configured to confirm SVs found in at least two tools. This led to creation of four merged VCF datasets [see Figure 1 for Data Set (DS) 1] for mule2mule, mule2white, white2white, and white2mule. Using BCFtools (Danecek et al., 2021), we filtered the structural variants to extract specific information for comparison, such as the number of SVs, their sizes in base pairs (bp), scores, and zygosity status.

We categorized each individuals SV based on whether it was detected in both the WTD and MD reference genomes or if it was specific to the reference assembly. To achieve this, we used Mummer (Delcher et al., 2002) to map the reference genomes of white-tailed deer (JAJQKH000000000) and mule deer (JAJLRB000000000). Utilizing the show-coords utility, we identified the coordinates of regions common between the two species. Each SV was then compared to these coordinates using a custom bash script,

resulting in the separation of SVs into two distinct files: one with unique SVs (only detected in one reference assembly) and another with shared SVs (detected in both reference assemblies). Lastly, we identified those SV present exclusively in one species (considered a species-specific SV); here we required SVs to be present in both 80% and 100% of samples analyzed in each species. This led to creation of two VCF datasets for WTD and MD species-specific SVs [see Figure 1 for DS2]. Lastly, we quantified the species-specific overlap of SVs between genome and SR data analysis streams.

2.3. Gene Enrichment and Selection Analysis and Enhancer Motifs

The resulting impact of SVs from both genome assembly and SR analysis was annotated using SnpEff (Cingolani et al., 2012). To enable this, we first created custom SnpEff databases using the genome FASTA files and their corresponding annotation GFF files. We used the genome assembly JAJLRB000000000 for mule deer and its associated annotation file from the NCBI database. We used the genome JAJQKH000000000 for white-tailed deer. To transfer annotations between assemblies we used Mummer (Delcher et al., 2002) and custom Python scripts. We then used UniProt and NCBI databases to retrieve detailed information on impacted genes. We used DAVID (The Database for Annotation, Visualization, and Integrated Discovery; Sherman et al., 2024) to analyze each gene set and obtain a comprehensive functional profile.

To calculate the ratio of nonsynonymous to synonymous substitution rates (dN/dS) for SV-specific genes identified in DS2, we collated additional genomes form Caribou (*Rangifer tarandus*, CATKSN000000000), Elk (*Cervus canadensis*, JACTMD000000000), Moose (*Alces alces*, CAXWQY000000000), and domestic cattle (*Bos taurus*, NKLS00000000). After identifying homologous gene sequences in each species using BLAST (Camacho et al., 2008) the nucleotide sequences were aligned using MAFFT (Katoh et al., 2005) and converted to a codon alignment using PRANK (Löytynoja, 2014). The aligned codon sequences, along with a user-specified species tree (cow, (elk, (caribou, ((mule, whitetailed), moose)))), were analyzed using the adaptive Branch-Site Random Effects Likelihood (aBSREL) (Smith et al., 2015) method implemented in HyPHy (Kosakovsky et al., 2005). Only genes

exhibiting statistically significant evidence of selection (dN/dS > 1 or dN/dS < 1, and p-value < 0.05) were retained for further interpretation.

We also screened WTD and MD genomes for enhancer motifs using Homer (Heinz et al., 2010). The findMotifsGenome, annotatePeaks, and scanMotifGenomeWide tools were employed and filtered based on the average motif scores. We extracted the number of motifs detected in intergenic regions and species-specific SVs identified in DS2; here, we normalized motif density, by dividing the number of motifs by the length of each region to produce a value per kilobase pair (kbp). To compare motifs between WTD and MD, we used only the WTD reference genome (JAJQKH000000000) [See supplementary data, Table 5]. All the mapping and calculation in whole analysis process was done using Python programming language.

3. Results

3.1. Genome SV exploration

We screened two genome assemblies (one WTD and one MD) against the WTD reference for SVs (Figure 1) resulting in two streams referred to as White_SV and Mule_SV. The average quality scores for WTD and MD SVs were 211 and 287 respectively. In total we found 188 SVs for MD and 91 SVs for WTD. The breakdown of SVs by method and species are shown in Table 2. Among the 188 SVs identified in mule deer, 56% were deletions, 1.5% were duplications, 36.7% were insertions, and 5.31% were translocations. For white-tailed deer 73.62% were deletions, 2.1% duplications, 23.07% insertions, and 1.09% translocations. Most deletion SVs were homozygous, and the majority of these SVs were intergenic.

3.2. Short read exploration (DS1)

The average quality score for SVs identified by Delly was 232, by Lumpy 171, and by Manta 218. The breakdown of SVs by method and species are shown in Table 3. Across all individuals in the merged SV file (i.e. 2/3 methods detected the SV); 59.81% were deletions, 0.76% duplications, 28.60% insertions,

1.30% inversions, and 9.52% were translocations (see Table 3 for complete summary). Interspecies mapped samples (i.e. mule2white) exhibited 14.99% more SVs compared to intraspecies samples (i.e. mule2mule), but the size and relative patterns were consistent across both interspecies and intraspecies mapped samples (Table 3). The majority of SVs were intergenic, though some impacted functional regions (Table 3).

3.3. Short read exploration (DS2)

We next focused on species specific SVs found in the majority of samples in each species (80% and 100%), with the 80% samples allowing for missing some SVs due to data quality or genome sampling. We identified 89 SVs only found in MD and 78 SVs only found in WTD. Out of 89 mule deer SVs, 13 were found in 100% of samples. Out of 78 white tailed deer SVs, 5 were found in 100% of samples. Some SVs were sufficiently large to impact multiple genes, leading to the identification of 143 impacted genes in mule deer and 127 impacted genes white-tailed deer. More information about these SVs and genes can be found in Table 4 and all of the genes and their functions are listed in supplementary data, Table 1-4.

3.3. Genome and enhancer impact

We performed both individual and gene set enrichment analysis to assess their potential impact and grouped them based on their functional similarities. This analysis revealed seven distinct functional clusters in mule deer and nine clusters in WTD, highlighting differences in their gene functions (Supplementary data Table 7-8). After masking genes and SVs associated with WTD and MD, the average motif density was found to be 5.04 per kilobase pair. Within WTD SVs, the average motif density was 5.86 motifs per kilobase pair, while MD SVs exhibited an average of 3.65 motifs per kilobase pair [See Supplementary data, Table 6]. The standard deviation of motif density in SVs for MD and WTD was 3.095 and 7.565, respectively.

We assessed selective pressures on SVs identified in the DS2 dataset of MD and white-tailed WTD. We analyzed 119 MD SV-associated genes and 121 WTD SV-associated genes for evidence of selection. Our

analysis revealed that 5 out of the 119 MD genes exhibited evidence of positive selection, characterized by a dN/dS ratio > 1 and a p-value < 0.05. Similarly, 4 out of the 121 WTD genes showed positive selection under the same criteria (Table 4, [Supplementary data Table 1-4]). A list of impacted genes and functions are included in Supplemental Table 1-4.

4. Discussion

The genomes of recent and hybridizing species provide a window into the mechanisms that drive biodiversity, adaptation, and reproductive isolation. We compared genomes on two closely related and hybridizing deer species, the white-tailed deer (*Odocoileus virginianus*) and mule deer (*O. hemionus*), to identify potential SVs involved in the speciation process. By integrating data from multiple genome assemblies and short-read sequencing, we identified SVs within and between these species, their impact on genes, and relationship to enhancers. The findings of species-specific SVs show how they underly genomic divergence potentially leading to reproductive isolation, reinforcing their role in shaping important evolutionary processes.

Characteristics of SVs in *Odocoileus*

Our analysis revealed a predominance of deletions among SVs in deer, followed by insertions and translocations. This finding is consistent with observations reported in other studies (Shi et al., 2023; Zhang et al., 2023), notably, Shi et al. (2023) identified deletions as the most frequent SV type in human populations. This dominance of deletions could be attributed to various mechanisms, including replication errors or unequal homologous recombination (Wold et al., 2023). The high proportion of homozygous deletions and insertions in our study points to these SVs being more likely to become fixed within a population compared to duplications and translocations that were more heterozygous (Table 3). Simmons et al. (2012) showed homozygous deletions often exhibit high penetrance; furthermore, the potential for homozygous deletions and insertions to stabilize in the genome could be linked to their lessened impact on recombination resulting in their tolerance and increased frequency. Berdan et al., (2024) showed that

SVs, particularly when heterozygous, can suppress recombination, but homozygous deletions and insertions, by their nature, do not inherently disrupt the linear arrangement of genetic material, minimizing their impact on recombination. Duplications and translocations, particularly in their homozygous form, can result in complex genomic rearrangements that disrupt gene order and heighten the risk of recombination errors. Zhang et al. (2021) discussed how gene duplications can drive speciation by disrupting gene function or fostering novel interactions; however, of the species-specific SVs, only 16% were duplications (Table 4), while the SV impacted genes with positive dN\dS values (9 total), were all either deletion or insertions. Here we note a limitation in evaluating the evolutionary consequences of duplications lies in SV detection, as current short-read-based tools identify but do not reconstruct the SV, with the absence of the SV in the assembly reflective of the common practice of purging duplications (Guan et al, 2020). As a result, dN/dS calculations are confined to the sequence of the initial (homologous) duplicated region, and we are currently unable to reliably assess the evolution of duplicated regions without long-read approaches.

Most of the SV impacts we observed were intergenic, and similar trends were reported by Yang et al. (2024) and David et al. (2024). These studies, like ours, demonstrated that a significant portion of SVs localize to intergenic regions in sheep, goats, and wild bird populations. This dominance of intergenic impacts can be attributed to the vast amount of non-coding DNA within the genome; however, these regions also contain important regulatory elements, which comprise a significant portion of the mammalian genome (Villar et al, 2015). We observed no differences in enhancer motif number (Supplementary data, Table 6), but mutations within conserve non-coding regions have been linked to various diseases, suggesting roles in gene regulation (Derežanin et al., 2023). David et al. (2024) observed that a high percentage of deletions and inversions in intergenic regions were classified as 'modifiers,' suggesting they could have phenotypic effects. Our findings provide a similar perspective where 74.21% of deletions and 52.94% of inversions occurred in intergenic regions and ff these, 34.43% of intergenic deletions and 23.27% of intergenic inversions were annotated as modifiers, reiterating the potential

consequential impacts SVs can have outside genic regions. Lastly, our analysis revealed a large abundance of SVs shared between mule deer and white-tailed deer. The prevalence of segregating SVs, most of which would be ancestral, highlights the conservation of genomic elements and potentially common evolutionary pressures acting on both species. Recent studies in Eucalyptus species (Ferguson et al., 2024) showed the role of shared SVs in adaptation and speciation with significant number shared SV polymorphisms being enriched for genes involved in adaptive traits, and thus further functional exploration and annotation assessments might reveal similar impacts in deer.

Species-species SVs, adaptation and potential incompatibilities

Several genes associated with species-specific SVs exhibited evidence of positive selection in both MD and WTD. These positively selected genes appear to be involved in key cellular and physiological processes that might lead to species-specific adaptations. In MD, the identified genes played a role in cellular differentiation, gene regulation, protein interactions, and metabolic processes, which could influence tissue development, immune responses, and overall cellular function, consistent with what Kessler et al. (2023) observed in deer. These findings also align with previous studies demonstrating that SVs can drive evolutionary divergence by altering gene function rather than enhancer activity (Quan et al., 2021). Structural variants have been shown to impact gene expression through changes in chromatin organization and regulatory interactions, potentially facilitating adaptive traits without necessarily modifying enhancer density (Razin et al., 2023); this is broadly consistent with our observation of no enhancer bias in SVs.

Several gene impacts were identified that align with known aspects of deer biology and speciation (Kessler et al., 2024). Notably, a MD specific deletion impacted gene A0A6J0YMR, which is involved in in male gonad development and sperm production. Kessler et. al. (2023) observed similar genes under selection in deer, and hybrid deer show evidence of reduced sperm function (Derr et al., 1991). Similarly, both Bracewell et al. and Dowle et al. (2017) showed how SVs on neo-sex chromosomes drive hybrid male sterility, demonstrating that chromosomal rearrangements can underpin the genetic basis of speciation via

genes directly involved in fertility. We also observed a WTD specific deletion impacting gene A0A6J0ZAC3 which is linked to olfactory function, similar processes also were identified in the SNP based approach of Kessler et al. (2024). Differences in olfactory genes might facilitate ecological partitioning between WTD and MD, contributing to their speciation. The role of olfactory genes in mate recognition and behavioral adaptation has been seen in house mice (North et al., 2020), where copy number variations in olfactory receptor genes influenced mate choice and reinforced reproductive isolation. This is particularly relevant in deer, as differences in olfactory responses may not only dictate ecological adaptations but also impact mate recognition, creating a dual mechanism for speciation through both environmental and sexual selection. Collectively, such SV alleles and the corresponding phenotypic variation might also implicate possible Dobzhansky-Muller incompatibilities in maintaining species boundaries in deer.

Differences between SV detection methods

The accurate estimation of SV size can be influenced by the specific tools and methodologies employed (Sedlazeck et al., 2018; Jeffares et al., 2017). The accuracy and sensitivity of SV callers vary significantly, as some tools are better suited for detecting larger SVs, while others excel at identifying smaller changes (Sedlazeck et al., 2018; Sudmant et al., 2015; Jeffares et al., 2017). This inherent bias across tools shapes the type of SV and the observed size distribution. To account for this, we employed two distinct data streams for SV detection to evaluate genome differences between MD and WTD, both reflective of commonly available datasets. The results from our direct genome comparisons deviated from those of previous studies that utilized long-read sequencing data (Shi et al., 2023; Sedlazeck et al., 2018). While the SV type ratio and size distribution of the detected SVs were generally consistent with expectations from prior long-read-based studies, the overall number of SVs identified was lower than expected. Although our genomes incorporated long-read data (Table 1), assembly artifacts may have reduced SV detection sensitivity. Furthermore, the relatively small sample size in our study—compared to, for example, the 320 individuals analyzed by Shi et al. (2023) likely led to the reduced detection in the genome

data and overlap with short-reads (DS1 and DS2). Thus, sample size is a paramount consideration in SV detection. Here, we utilized multiple references to alleviate potential bias (Yang et al, 2020) and focussed on unique but shared SVs within species to increase confidence in SV detection (Figure 1); this workflow should be accessible to most wildlife genome data sets.

Acknowledgements: We thank Alan Cain, Steeve Côté, Anh Dao, Marco Festa-Bianchet, Brad Fulk, Eric Hoffman, Levi Jaster, Daniel Koelsch, Emily Latch, Brent Patterson, Joe Nocera, Don Stewart, Russell Easy, Jonathan Shaw, David Walter, and Jon Wheeler for providing samples.

Data availability: All short-read sequencing samples are publicly available in the NCBI database under BioProject accession PRJNA830519. The code supporting this work is available in our lab's GitLab repository:

https://gitlab.com/WiDGeT_TrentU/graduate_theses/-/tree/master/AzimiChetabi/Chapter%20one

Author Contribution: We had no external contributions.

Funding: We were supported by Trent University's International Graduate Scholarship, an Ontario Graduate Scholarship, and Dean's PhD Scholarship. This work was supported by the Natural Sciences and Engineering Research Council of Canada Discovery Grant (grant number: RGPIN-2017-03934), ComputeCanada Resources for Research Groups (grant number: RRG gme-665-ab), Canada Foundation for Innovation: John R. Evans Leaders Fund, and Ontario Early Researcher Award (grant number: #36905).

Conflict of interest statement: The authors declare no competing interests.

References

Abril, V. V., E. A. G. Carnelossi, S. González, and J. M. B. Duarte. 2010. Elucidating the evolution of the red brocket deer Mazama americana complex (Artiodactyla; Cervidae). Cytogenetic and Genome Research 128:177–187.

Baurens, F.-C., G. Martin, C. Hervouet, et al. 2019. Recombination and large structural variations shape interspecific edible bananas genomes. *Molecular Biology and Evolution* 36(1):97–111.

Berdan, E. L., T. G. Aubier, S. Cozzolino, R. Faria, J. L. Feder, M. D. Giménez, M. Joron, J. B. Searle, and C. Mérot. 2024. Structural Variants and Speciation: Multiple Processes at Play. Cold Spring Harb Perspect Biol. 16:a041446.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114-20. doi: 10.1093/bioinformatics/btu170. Epub 2014 Apr 1. PMID: 24695404; PMCID: PMC4103590.

Broad Institute. 2018. Picard Toolkit [GitHub repository]. Retrieved from http://broadinstitute.github.io/picard/

Bracewell, R. R., B. J. Bentz, B. T. Sullivan, and J. M. Good. 2017. Rapid Neo-Sex Chromosome Evolution and Incipient Speciation in a Major Forest Pest. Nat. Commun. 8:1593.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinformatics. 2009 Dec 15;10:421. doi: 10.1186/1471-2105-10-421. PMID: 20003500; PMCID: PMC2803857.

Chen, X., O. Schulz-Trieglaff, R. Shaw, B. Barnes, F. Schlesinger, M. Källberg, et al. 2016. Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics 32:1220–1222.

Cingolani, P., A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, and D. M. Ruden. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly 6:80–92.

Cursino, M. S., M. B. Salviano, V. V. Abril, et al. 2014. The role of chromosome variation in the speciation of the red brocket deer complex: The study of reproductive isolation in females. BMC Evolutionary Biology 14:40.

Danecek, P., J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, and H. Li. 2021. Twelve years of SAMtools and BCFtools. GigaScience 10:giab008.

David, G., A. Bertolotti, R. Layer, D. Scofield, A. Hayward, T. Baril, H. A. Burnett, E. Gudmunds, H. Jensen, and A. Husby. 2024. Calling structural variants with confidence from short-read data in wild bird populations. Genome Biology and Evolution 16:evae049.

Delcher, A. L., A. Phillippy, J. Carlton, and S. L. Salzberg. 2002. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Research 30:2478–2483.

Derežanin, L., J. Fickel, R. Tiedemann, and P. C. Alves. 2023. Contribution of structural variation to adaptive evolution of mammalian genomes. Universität Potsdam.

Derr, J. N., D. W. Hale, D. L. Ellsworth, and J. W. Bickham. 1991. Fertility in an F1 male hybrid of white-tailed deer (Odocoileus virginianus) x mule deer (O. hemionus). J Reprod Fertil. 93:111-7.

Dowle, E. J., R. R. Bracewell, M. E. Pfrender, K. E. Mock, B. J. Bentz, and G. J. Ragland. 2017. Reproductive Isolation and Environmental Adaptation Shape the Phylogeography of Mountain Pine Beetle (Dendroctonus Ponderosae). Mol. Ecol. 26:6071–6084.

Escaramís, G., E. Docampo, and R. Rabionet. 2015. A decade of structural variants: Description, history, and methods to detect structural variation. Briefings in Functional Genomics 14:367–379.

Ferguson, S., A. Jones, K. Murray, R. L. Andrew, B. Schwessinger, H. Bothwell, J. Borevitz, 2024. Exploring the role of polymorphic interspecies structural variants in reproductive isolation and adaptive divergence in Eucalyptus. GigaScience 13:giae029.

Gompert, Z., J. L. Feder, T. L. Parchman, et al. 2025. Adaptation repeatedly uses complex structural genomic variation. *Science* 388(6744):eadp3745.

Gonzalez, S., and J. M. B. Duarte. 2020. Speciation, evolutionary history and conservation trends of Neotropical deer. *Mastozoología Neotropical* 27(SI):35–46.

Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics. 2020 May 1;36(9):2896-2898. doi: 10.1093/bioinformatics/btaa025. PMID: 31971576; PMCID: PMC7203741.

Heinz, S., C. Benner, N. Spann, E. Bertolino et al. 2010. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell 38:576-589.

Homolka, D., R. Ivánek, J. Capkova, P. Jansa, and J. Forejt. 2007. Chromosomal rearrangement interferes with meiotic X chromosome inactivation. Genome Research 17:1431–1437.

Jeffares, D. C., C. Jolly, M. Hoti, D. Speed, L. Shaw, C. Rallis, et al. 2017. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. Nature Communications 8:14061.

Kaplow, I. M., M. T. Maurano, A. Kundaje, et al. 2023. Relating enhancer genetic variation across mammals to complex phenotypes using machine learning. *Science* 380(6648):eabm7993.

Katoh, K., and D. M. Standley. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution* 30(4):772–780.

Kess, T., A. L. Einfeldt, B. Wringe, S. J. Lehnert, K. K. S. Layton, M. C. McBride, D. Robert, J. Fisher, A. Le Bris, C. den Heyer, N. Shackell, D. E. Ruzzante, P. Bentzen, and I. R. Bradbury. 2021. A putative structural variant and environmental variation associated with genomic divergence across the Northwest Atlantic in Atlantic Halibut. ICES Journal of Marine Science 78:2371–2384.

Kessler, C., and A. B. A. Shafer. 2023. Genomic analyses capture the human-induced demographic collapse and recovery in a wide-ranging cervid. bioRxiv. https://doi.org/10.1101/2023.07.19.

Kessler, C., E. Wootton, and A. B. A. Shafer. 2023. Speciation without gene-flow in hybridizing deer. Mol Ecol. 32:1117-1132.

Kosakovsky Pond, S. L., S. D. W. Frost, and S. V. Muse. 2005. HyPhy: Hypothesis testing using phylogenies. *Bioinformatics* 21(5):676–679.

Langmead, B., and S. L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods 9:357–359.

Layer, R. M., C. Chiang, A. R. Quinlan, and I. M. Hall. 2014. LUMPY: A probabilistic framework for structural variant discovery. Genome Biology 15:R84.

Li, H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100.

London, E. W., A. L. Roca, J. E. Novakofski, and N. E. Mateus-Pinilla. 2022. A De Novo Chromosome-Level Genome Assembly of the White-Tailed Deer, Odocoileus Virginianus. Journal of Heredity 113:479–489.

Mahmoud, M., N. Gobet, D. I. Cruz-Dávalos, et al. 2019. Structural variant calling: the long and the short of it. Genome Biology 20:246.

Martin, G., F. Carreel, O. Coriton, C. Hervouet, C. Cardi, P. Derouault, D. Roques, F. Salmon, M. Rouard, J. Sardos, K. Labadie, F. C. Baurens, and A. D'Hont. 2017. Evolution of the banana genome (Musa acuminata) is impacted by large chromosomal translocations. Molecular Biology and Evolution 34:2140–2152.

North, H. L., P. Caminade, D. Severac, K. Belkhir, and C. M. Smadja. 2020. The role of copy-number variation in the reinforcement of sexual isolation between the two European subspecies of the house mouse. Phil. Trans. R. Soc. B 375:20190540.

Löytynoja A. Phylogeny-aware alignment with PRANK. 2014. Methods Mol Biol. 2014;1079:155-70. doi: 10.1007/978-1-62703-646-7_10. PMID: 24170401.

Pavlidis, P., J. D. Jensen, W. Stephan, and A. Stamatakis. 2012. A Critical Assessment of Storytelling: Gene Ontology Categories and the Importance of Validating Genomic Scans. Molecular Biology and Evolution 29:3237–3248.

Pertea, M., G. M. Pertea, C. M. Antonescu, T.-C. Chang, J. T. Mendell, and S. L. Salzberg. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature Biotechnology 33:290–295.

Quan C, Li Y, Liu X, Wang Y, Ping J, Lu Y, Zhou G. Characterization of structural variation in Tibetans reveals new evidence of high-altitude adaptation and introgression. Genome Biol. 2021 May 25;22(1):159. doi: 10.1186/s13059-021-02382-3. PMID: 34034800; PMCID: PMC8146648.

Quinlan, A. R., and I. M. Hall. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842.

Rausch, T., T. Zichner, A. Schlattl, A. M. Stutz, V. Benes, and J. O. Korbel. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28:i333–i339.

Razin, S. V., S. V. Ulianov, and O. V. Iarovaia. 2023. Enhancer function in the 3D genome. *Genes* 14(6):1277.

Sedlazeck, F. J., P. Rescheneder, M. Smolka, et al. 2018. Accurate detection of complex structural variations using single-molecule sequencing. Nature Methods 15:461–468.

Shang, H., D. L. Field, O. Paun, M. Rendón-Anaya, J. Hess, C. Vogl, J. Liu, P. K. Ingvarsson, C. Lexer, and T. Leroy. 2023. Drivers of genomic landscapes of differentiation across a Populus divergence gradient. Molecular Ecology 32:4348–4361.

Sherman, B. T., G. Panzade, T. Imamichi, and W. Chang. 2024. DAVID ortholog: An integrative tool to enhance functional analysis through orthologs. Bioinformatics. Advance online publication.

Shi, J., Z. Jia, J. Sun, et al. 2023. Structural variants involved in high-altitude adaptation detected using single-molecule long-read sequencing. Nat Commun 14:8282.

Simmons, A. D., C. M. Carvalho, and J. R. Lupski. 2012. What have studies of genomic disorders taught us about our genome? Methods Mol Biol 838:1-27.

Smith, M. D., J. O. Wertheim, S. Weaver, et al. 2015. Less is more: An adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Molecular Biology and Evolution* 32(5):1342–1353.

Sobel, J. M., G. F. Chen, L. R. Watt, and D. W. Schemske. 2010. The biology of speciation. *Evolution* 64:295–315. https://doi.org/10.1111/j.1558-5646.2009.00877.x

SRA Toolkit Development Team. 2024. SRA Toolkit. Available from: https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software

Stevison, L. S., K. B. Hoehn, and M. A. Noor. 2011. Effects of inversions on within- and between-species recombination and divergence. *Genome Biology and Evolution* 3:830–841. https://doi.org/10.1093/gbe/evr081

Sudmant, P. H., T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* 526:75–81.

Villar, D., C. Berthelot, S. Aldridge, et al. 2015. Enhancer evolution across 20 mammalian species. *Cell* 160(3):554–566.

Wold, J. R., J. G. Guhlin, P. K. Dearden, A. W. Santure, and T. E. Steeves. 2023. The promise and challenges of characterizing genome-wide structural variants: A case study in a critically endangered parrot. Mol Ecol Resour.

Yang L. A Practical Guide for Structural Variation Detection in the Human Genome. Curr Protoc Hum Genet. 2020 Sep;107(1):e103. doi: 10.1002/cphg.103. PMID: 32813322; PMCID: PMC7738216.

Yang, J., D. F. Wang, J. H. Huang, et al. 2024. Structural variant landscapes reveal convergent signatures of evolution in sheep and goats. Genome Biology 25:148.

Zhang, L., S. Chaturvedi, C. C. Nice, L. K. Lucas, and Z. Gompert. 2023. Population genomic evidence of selection on structural variants in a natural hybrid

Figure caption:

Figure 1. Workflow used in our study. The map at the top shows the geographical range of the two species under investigation. The light grey area represents the range of white-tailed deer, while the grey area indicates the range of mule deer. The dark grey region highlights the overlap between the two species' habitats. Each point indicates the specific locations where samples were collected. Below the map, the workflow is divided into two distinct pipelines: one for short-read analysis and the other for genome data analysis.

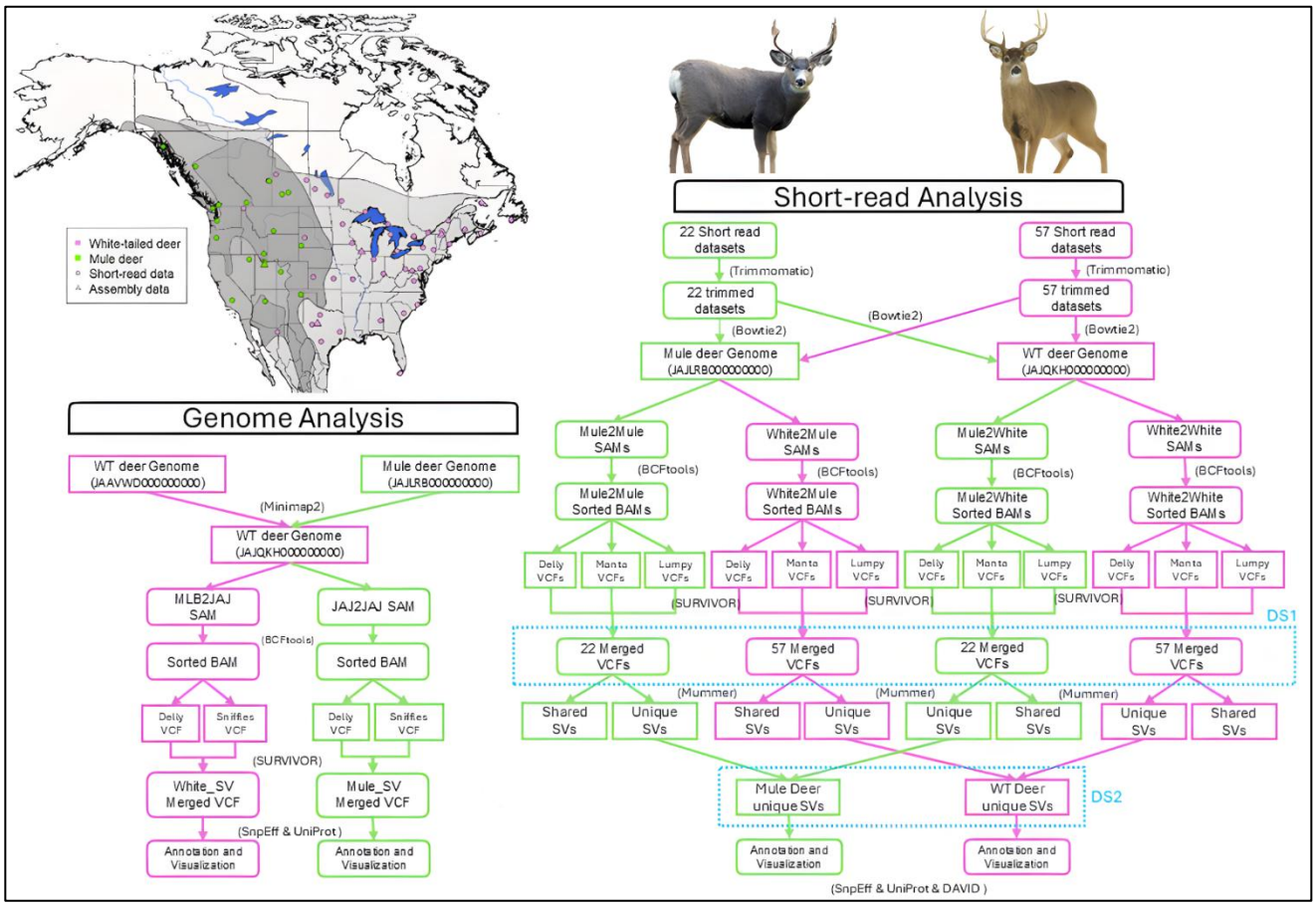

Table caption:

Table 1. White-tailed deer and Mule deer genome that were used in genome data and short-read analysis. Sequencing used in genome assembly was long read (LR) or short read (SR) data.

Species	*Odocoileus virginianus*	*Odocoileus virginianus*	*Odocoileus hemionus*
Bio-project	PRJNA782613	PRJNA420098	PRJNA752226
Access NO.	JAJQKH000000000	JAAVWD000000000	JAJLRB000000000
Citation	(London et al,2022)	(Anderson et al, 2022)	(Lamb et al, 2021)
Sequencing Used	SR	LR,SR	LR,SR
Year	2022	2020	2021
Genome Size	2.42 Gb	2.5Gb	2.6 Gb
# Scaffolds	480	2905	5513
Scaffold N50	66.41 Mb	17.1 Mb	72.1 Mb

Table 2. General structural variant (SV) statistics (Type of SV, Total number of detected SVs, Homozygosity Percentage, Impact level, Species) for genome assembly analysis for both mule deer (Mule_SV) and White-tailed deer (White_SV).

TYPE	NO.	SIZE (AVG)	HOMO (%)	EXON/INTRON/GENE-LEVEL/INTERGENIC	REF
DELETION	106	521,068	81.13%	6 / 22 /2 / 76	Mule_SV
DUPLICATION	3	27,567,413	0%	1 / 0 / 0 /2	Mule_SV
INSERTION	69	20,626	59.42%	1 / 18 / 0 / 50	Mule_SV
INVERSION	0	0	0%	0	Mule_SV
TRANSLOCATION	10	8,013,211	20%	0 / 0/ 1 / 9	Mule_SV
DELETION	67	210,234	55.22%	6 / 18 / 1 / 42	White_SV
DUPLICATION	2	15,701,375	50%	1 / 0 / 1 / 0	White_SV
INSERTION	21	20,600	57.14%	0 / 6 / 0 / 15	White_SV
INVERSION	0	0	0	0	White_SV
TRANSLOCATION	1	7,265,910	0%	0 / 0 / 1 / 0	White_SV

Table 3. General structural variant (SV) statistics (Type of SV, Total number of detected SVs, Homozygosity Percentage, Impact level, Dataset) for dataset 1 (DS1) analysis.

TYPE	NO.	SIZE (AVG)	HOMO (%)	EXON/INTRON/GENE-LEVEL/INTERGENIC	REF
DELETION	38,612	21,257	75.47%	1,086 / 5,385 / 1,179 / 30,962	M2M
DUPLICATION	553	2,558,746	12.16%	80 /88 / 78 / 307	M2M
INSERTION	1,7834	25,235	58.05%	662 /2,768 / 487 / 13,917	M2M
INVERSION	463	7,227,436	49.18%	1 / 16 / 185 / 261	M2M
TRANSLOCATION	4,499	6,410,309	23.70%	607 / 595 / 1,914 / 1,383	M2M
DELETION	107,684	24,813	72.82%	6,271 /28,158 /3,367 / 69,888	M2W
DUPLICATION	1,582	2,063,234	7.46%	259 / 333/ 211 / 779	M2W
INSERTION	60,303	24,395	91.68%	2,826 / 16,567 / 2,356/ 38,554	M2W
INVERSION	2,749	3,425,350	30.12%	182 / 448 / 814 / 1,305	M2W
TRANSLOCATION	23,308	6,918,217	19.83%	2,745 / 2,486 / 11,260 / 6,817	M2W
DELETION	239,789	12,080	74.14%	3,767 / 35,094 / 5,133 / 195,795	W2M
DUPLICATION	1,956	5,670,389	9.81%	225 / 325 / 211 / 1,195	W2M
INSERTION	122,788	26,139	74.86%	1,572 / 20,155 / 1,273 / 99,788	W2M
INVERSION	4,776	5,913,090	39.65%	106 / 394 / 1,248 / 3,028	W2M
TRANSLOCATION	31,060	6,180,882	21.76%	1,663 / 1,557 / 15,748/ 12,092	W2M
DELETION	163,143	856,900	64.79%	7,421 / 42,635 / 2,238 / 110,849	W2W
DUPLICATION	2,887	1,209,213	5.34%	466 / 606 / 332 / 1,483	W2W
INSERTION	61,649	25,013	88.94%	806 / 20,276 / 427 / 40,140	W2W
INVERSION	3,959	2,176,793	25.51%	377 / 822 / 1,029 / 1,731	W2W
TRANSLOCATION	28,543	6,683,432	21.01%	31,109 / 2,773 / 14,009 / 8,652	W2W

Table 4. General information about genes found only in mule deer or white-tailed deer (dataset 2 (DS2), Genome data.). The number in parentheses next to number of genes in 80% shows the number of significant genes with dN/dS>1 or dN/dS<1, and PValue <0.05.

	Mule deer		
	DS2		**Genome Data**
	100% of samples	**80% of samples**	**All samples**
Number of SVs	13	76	67
SV type	Del:9 Inv:2 Tra:2	Del:38 Dup:3 Inv:4 Ins:27 Tra:3	Del:58 Dup:2 Ins:2 Tra:5
Size Avg (bp)	Del: 24,257 Inv: 4,586,383 Tra: 2,921,768	Del:1229 Dup:25,597 Ins:20,315 Tra:2,080,262	Del:1,279 Dup: 8413 Ins:3,024 Tra: 306,339
# Genes	24	119 (5)	122
Gene Impact	Exon:2 Intron:3 Gene:7 Inter:12	Exon:14 Intron:21 Gene:17 Inter:67	Gene:4 Inter:85
	White-tailed deer		
	DS2		**Genome Data**
	100% of samples	**80% of samples**	**All samples**
Number of SVs	5	73	19
SV type	Del:3 Ins:1 Tra:1	Del:51 Inv:3 Ins: 16 Tra:1	Del:18 Dup:1 Ins:1
Size Avg (bp)	Del:7306 Ins: 552 Tra:1,265,408	Del:1760 Inv:22,609 Ins:1863 Tra: 2,802,690	Del: 11,237 Dup: 1,701,375 Ins: 17,231
# Genes	6	121 (4)	78
Gene Impact	Exon:1 Intron:1 Gene:2 Inter:2	Exon:2 Intron:16 Gene:7 Inter:58	Exon:7 Intron:20 Gene:2 Inter:49